FEATURE ENGINEERING: MUHIM HUSUSIYATLARNI TANLASH VA YARATISH USULLARI

Olimjonova Havasxon Orzubek qizi
FarDU Axborot tizimlari va texnologiyalari yo‘nalishi 3-kurs talabasi.
havasxonolimjonova349@gmail.com
Sobirjonov Behzod Qahramonovich
FarDU Axborot texnologiyalari kafedrasi o‘qituvchisi
bekzodbekqahromonovich@gmail.com
https://doi.org/10.5281/zenodo.20354979

Anotatsiya: Maqolada mashina o‘rganish modellarining samaradorligini oshirishda Feature Engineering (xususiyatlar muhandisligi) jarayonining o‘rni va ahamiyati yoritilgan. Unda xususiyatlarni tanlashning statistik (filtrlash), iteratsion (o‘rash) va model ichiga o‘rnatilgan usullari batafsil tahlil qilingan. Shuningdek, yangi xususiyatlar yaratish strategiyalari, ma’lumotlarni transformatsiya qilish, domen bilimining ahamiyati hamda Python ekotizimidagi asosiy kutubxonalar (Pandas, Scikit-learn) ko‘rib chiqilgan. Xulosada sohaning avtomatlashtirish istiqbollari va mutaxassis intuitsiyasining o‘rni ta’kidlangan.

Kalit so‘zlar: Feature Engineering, xususiyatlarni tanlash, xususiyatlarni yaratish, overfitting, ma’lumotlar transformatsiyasi, korrelyatsiya, Lasso, mashina o‘rganish, Pandas, Scikit-learn.

FEATURE ENGINEERING: HOW TO SELECT AND CREATE IMPORTANT FEATURES

Olimjonova Havaskhon Orzubek kizi
3rd-year student at FarDU majoring in Information Systems and Technologies.
havasxonolimjonova349@gmail.com
Sobirjonov Bekhzod Kakhramonovich
Lecturer of the Department of Information Technologies at FarDU
bekzodbekqahromonovich@gmail.com

Annotation: The article highlights the role and significance of the Feature Engineering process in enhancing the efficiency of machine learning models. It provides a detailed analysis of feature selection methods, including statistical (filtering), iterative (wrapper), and embedded techniques. Additionally, strategies for creating new features, data transformation, the importance of domain knowledge, and key libraries in the Python ecosystem (Pandas, Scikit-learn) are discussed. The conclusion emphasizes the prospects of automation in the field and the crucial role of expert intuition.

Keywords: Feature Engineering, feature selection, feature creation, overfitting, data transformation, correlation, Lasso, machine learning, Pandas, Scikit-learn.

ИНЖЕНЕРИЯ ФУНКЦИЙ: СПОСОБЫ ВЫБОРА И СОЗДАНИЯ ВАЖНЫХ ФУНКЦИЙ

Олимжонова Хавасхон Орзубек кизи
Студентка 3 курса направления Информационные системы и технологии ФерГУ.
havasxonolimjonova349@gmail.com

Собиржонов Бехзод Кахрамонович
Преподаватель кафедры информационных технологий ФерГУ
bekzodbekqahromonovich@gmail.com

Аннотация: В статье освещаются роль и значение процесса Feature Engineering (проектирование признаков) в повышении эффективности моделей машинного обучения. Подробно анализируются статистические (фильтрация), итерационные (обертка) и встроенные методы отбора признаков. Также рассматриваются стратегии создания новых признаков, трансформация данных, важность доменных знаний и основные библиотеки в экосистеме Python (Pandas, Scikit-learn). В заключении подчеркиваются перспективы автоматизации отрасли и роль интуиции специалиста.

Ключевые слова: Feature Engineering, отбор признаков, создание признаков, переобучение, трансформация данных, корреляция, Lasso, машинное обучение, Pandas, Scikit-learn.

Feature Engineering, yoki xususiyatlarni yaratish va tanlash jarayoni, mashina o‘rganish modellarining samaradorligini sezilarli darajada oshirishga xizmat qiluvchi fundamental bosqich hisoblanadi. Bu jarayon xom ma’lumotlardan model uchun eng axborotli, aniq va foydali xususiyatlarni ajratib olish hamda ularni transformatsiya qilishni nazarda tutadi. Ma’lumotlarning sifatini va tuzilishini yaxshilash orqali, Feature Engineering modellarning o‘rganish qobiliyatini optimallashtiradi va hatto murakkab algoritmlar uchun ham yuqori bashorat aniqligini ta’minlaydi. Uning asosiy ahamiyati bir nechta jihatlarda namoyon bo‘ladi. Birinchidan, u ma’lumotlardagi yashirin naqshlarni va munosabatlarni aniqlashga yordam beradi. Misol uchun, vaqt qatorlari ma’lumotlarida kunning vaqtiga bog‘liq o‘zgarishlarni aks ettiruvchi yangi xususiyatlar (masalan, "kunning qaysi choragi") qo‘shish modelning samaradorligini oshirishi mumkin. Ikkinchidan, u ma’lumotlarning dimensionalityini kamaytirishga yordam beradi, bu esa modelni o‘qitish vaqtini qisqartiradi hamda overfitting (haddan tashqari moslashish) xavfini kamaytiradi. Uchinchidan, Feature Engineering modelning interpretatsiya qobiliyatini yaxshilaydi. Yaxshi yaratilgan xususiyatlar modelning qanday qaror qabul qilayotganini tushunishni osonlashtiradi. Bu jarayon ko‘pincha domenga oid chuqur bilimlarni talab qiladi. Masalan, moliyaviy ma’lumotlar bilan ishlaganda, "o‘rtacha harakatlanuvchi narx" yoki "volatillik" kabi xususiyatlarni yaratish ekspert bilimiga asoslanadi. Har bir dastur uchun eng mos xususiyatlarni tanlash va yaratish individual yondashuvni talab qiladi va modelning muvaffaqiyatida hal qiluvchi rol o‘ynaydi. Shunday qilib, Feature Engineering nafaqat texnik jarayon, balki ijodiy va strategik yondashuv hamdir.

Mavjud xususiyatlarni tanlash va ularni optimallashtirish usullari Mavjud xususiyatlarni tanlash va optimallashtirish mashinani o‘rganish modellarining samaradorligini oshirishda hal qiluvchi ahamiyatga ega. Bu jarayon modelning murakkabligini kamaytirish, hisoblash narxini pasaytirish va haddan tashqari moslashish (overfitting) xavfini minimallashtirishga yordam beradi. Asosiy maqsad eng informativ xususiyatlar to‘plamini aniqlashdir. Filtrlash usullari statistik mezonlardan foydalanib, xususiyatlarni mustaqil ravishda baholaydi. Masalan, Korrelyatsiya Koeffitsienti (Pearson, Spearman) xususiyatlar va maqsad o‘zgaruvchisi o‘rtasidagi chiziqli yoki monotonik bog‘liqlikni o‘lchaydi. Yuqori korrelyatsiyaga ega bo‘lgan xususiyatlar odatda saqlanadi. O‘zaro Ma’lumot (Mutual Information) esa chiziqli bo‘lmagan bog‘liqliklarni ham aniqlay oladi, bu esa ma’lumotlar to‘plamining murakkab tuzilmalarini tushunishda foydali. Chi-kvadrat (Chi-Squared) testi diskret xususiyatlar uchun maqsad o‘zgaruvchisi bilan bog‘liqlikni

baholashda keng qo‘llaniladi. Bu usullar tezkor va hisoblash jihatdan samarali bo‘lsa-da, xususiyatlar orasidagi o‘zaro ta’sirlarni hisobga olmaydi. O‘rash usullari (Wrapper Methods) esa tanlangan xususiyatlar to‘plamida modelni o‘qitish va uning ishlashini baholash orqali xususiyatlarni tanlashni amalga oshiradi. Masalan, Oldinga Tanlash (Forward Selection) bo‘sh to‘plamdan boshlab, eng yaxshi natija beradigan xususiyatni bosqichma-bosqich qo‘shib boradi. Orqaga Qaytarish (Backward Elimination) esa barcha xususiyatlar bilan boshlab, modelning ishlashiga eng kam ta’sir qiluvchi xususiyatni olib tashlab boradi. Rekursiv Xususiyatlarni Olib Tashlash (Recursive Feature Elimination, RFE) esa modelni o‘qitadi, xususiyatlar muhimligini aniqlaydi va eng kam muhimini olib tashlaydi, bu jarayonni takrorlaydi. Bu usullar modelning ishlashini to‘g‘ridan-to‘g‘ri optimallashtiradi, ammo hisoblash jihatdan qimmatroq bo‘ladi. O‘rnatilgan usullar (Embedded Methods) xususiyat tanlashni modelni o‘qitish jarayoniga birlashtiradi. Misol uchun, Lasso (L1 regularization) regressiyasi ba’zi xususiyatlarning koeffitsientlarini nolga tushirib, ularni samarali ravishda yo‘q qiladi. Qaror Daraxtlari (Decision Trees) va Tasodifiy O‘rmonlar (Random Forests) kabi modellar xususiyatlarning ahamiyatini avtomatik ravishda baholaydi va eng muhimlarini tanlashda yordam beradi. Gradient Boosting mashinalari ham xususiyat ahamiyatini aniqlash imkoniyatini beradi. Bu usullar filtrlash va o‘rash usullarining afzalliklarini birlashtirib, samarali va aniq natijalar beradi. Xususiyatlarni optimallashtirish nafaqat ularni tanlashni, balki ularni o‘zgartirishni ham o‘z ichiga oladi. Yangi xususiyatlar yaratish (feature creation) va mavjudlarini o‘zgartirish (transformation) ham modelning aniqligini oshirishi mumkin. Masalan, logarifmik transformatsiya qiymatlar diapazonini barqarorlashtirishi yoki polynomial transformatsiya chiziqli bo‘lmagan bog‘liqliklarni ifodalashi mumkin. Bu usullarning har biri o‘ziga xos afzallik va kamchiliklarga ega bo‘lib, eng yaxshi yondashuv ma’lumotlar to‘plami va modelning turiga qarab farqlanadi.

Yangi xususiyatlarni yaratish samarali model ishlashi uchun muhimdir. Bu jarayon mavjud ma’lumotlardan foydalanib, modelning tushunishini yaxshilaydigan yangi o‘zgaruvchilar yaratishni o‘z ichiga oladi. Asosiy strategiyalardan biri domen bilimini qo‘llashdir. Masalan, moliyaviy ma’lumotlarda, har bir mijozning qarzdorlik-daromad nisbatini hisoblash, kredit xavfini baholashda foydali bo‘lishi mumkin. Bu nisbat to‘g‘ridan-to‘g‘ri mavjud bo‘lmasa-da, "qarz" va "daromad" ustunlaridan olinadi. Boshqa bir kuchli yondashuv ma’lumotlarni o‘zgartirish orqali yangi xususiyatlar yaratishdir. Logaritmik o‘zgartirish qiyshiq taqsimlangan ma’lumotlarni normal holatga keltirishi mumkin, bu esa chiziqli modellar uchun foydalidir. Masalan, uy narxlari ma’lumotlarida log(narx)ni hisoblash modelning bashorat aniqligini oshirishi mumkin, chunki uy narxlari ko‘pincha o‘ng tomonga egilgan taqsimotga ega bo‘ladi. Polinomial xususiyatlar yaratish ham muhimdir, unda mavjud xususiyatlarning darajalari (masalan, x^2, x^3) yoki ularning o‘zaro ta’sirlari (masalan, x*y) qo‘shiladi. Bu, ayniqsa, chiziqli bo‘lmagan munosabatlarni aniqlashda yordam beradi. Misol uchun, reklama xarajatlari va sotuvlar o‘rtasida chiziqli bo‘lmagan bog‘liqlik bo‘lsa, reklama_xarajatlari^2 ni qo‘shish modelning bu munosabatni yaxshiroq o‘rganishiga imkon beradi. Vaqt seriyalari ma’lumotlarida kechikish xususiyatlarini yaratish (lag features) juda foydalidir. Bu, oldingi vaqt nuqtalaridagi qiymatlarni hozirgi vaqt nuqtasiga xususiyat sifatida qo‘shishni anglatadi. Masalan, ob-havoni bashorat qilishda, kechagi haroratni bugungi haroratni bashorat qilish uchun xususiyat sifatida ishlatish modelning samaradorligini oshirishi mumkin. Agregatsiya xususiyatlari ham keng qo‘llaniladi. Guruhdagi o‘rtacha qiymat, maksimum, minimum yoki standart og‘ish kabi statistik ko‘rsatkichlar foydali yangi xususiyatlar yaratishi mumkin. Masalan, mijozlarning xaridlar tarixi bo‘yicha ularning oxirgi uch oydagi o‘rtacha xarid summasini hisoblash, kelajakdagi xaridlar ehtimolini bashorat qilishda yordam beradi. Ushbu strategiyalar

ma’lumotlarning yashirin naqshlarini ochib berish va o‘rganish modelining kuchini oshirishga xizmat qiladi.

Mashina o‘rganish modellarining aniqligini oshirishda xususiyat muhandisligining hal qiluvchi rolini hisobga olsak, ushbu jarayonni soddalashtirish va optimallashtirish uchun turli xil vositalar va kutubxonalar ishlab chiqilgan. Bu vositalar ma’lumotlarni o‘zgartirish, yangi xususiyatlarni yaratish va mavjud xususiyatlarni tanlash bo‘yicha keng imkoniyatlarni taqdim etadi. Python ekotizimida Pandas kutubxonasi jadval ma’lumotlarini manipulyatsiya qilish uchun asosiy vosita bo‘lib xizmat qiladi. U DataFrame ob’ektlari orqali ma’lumotlarni import qilish, tozalash, birlashtirish va agregatsiya qilish uchun kuchli funksiyalarni taqdim etadi. Masalan, Pandas yordamida ustunlarni birlashtirish orqali yangi xususiyatlarni yaratish, kategoriya ma’lumotlarni One-Hot Encoding yordamida raqamli shaklga o‘tkazish yoki o‘rtacha qiymat bilan etishmayotgan qiymatlarni to‘ldirish mumkin. Scikit-learn (sklearn) kutubxonasi xususiyat muhandisligi uchun keng qamrovli vositalar to‘plamini o‘z ichiga oladi. U Preprocessing moduli orqali ma’lumotlarni standartlashtirish (StandardScaler), normalizatsiya qilish (MinMaxScaler) va kategoriya ma’lumotlarni kodlash (LabelEncoder, OneHotEncoder) kabi muhim funksiyalarni taklif etadi. Bundan tashqari, sklearn xususiyat tanlash usullarini ham o‘z ichiga oladi, masalan, SelectKBest yordamida statistik testlarga asoslangan eng yaxshi K xususiyatni tanlash yoki PCA (Principal Component Analysis) yordamida o‘lchovlilikni kamaytirish. Matn ma’lumotlari bilan ishlashda NLTK (Natural Language Toolkit) va SpaCy kutubxonalari muhim ahamiyatga ega. Ular so‘zlarni tokenizatsiya qilish, lemmatizatsiya, stop-so‘zlarni olib tashlash va TF-IDF (Term Frequency-Inverse Document Frequency) kabi matn xususiyatlarini yaratishda yordam beradi. Katta hajmdagi ma’lumotlar to‘plamlari uchun Dask va Spark kabi taqsimlangan hisoblash ramkalari xususiyat muhandisligi jarayonlarini parallel ravishda bajarish imkoniyatini beradi. Ular Pandas yoki Scikit-learnning APIlariga o‘xshash interfeyslarni taqdim etib, katta ma’lumotlar bilan ishlashni soddalashtiradi. Deep Learning modellarida xususiyat muhandisligi kamroq talab etilsa-da, TensorFlow va PyTorch kabi ramkalar maxsus qatlamlar orqali avtomatik ravishda xususiyatlarni o‘rganish imkoniyatini beradi. Masalan, konvolyutsion tarmoqlar tasvirlardan avtomatik tarzda ierarxik xususiyatlarni ajratib oladi. Ushbu vositalardan samarali foydalanish modelning ishlashini sezilarli darajada yaxshilashga va ma’lumotlar tahlilini tezlashtirishga yordam beradi.

XULOSA: Feature Engineeringning kelajagi va amaliy qo‘llanilishi bo‘yicha ushbu tahlilga yakun yasab, shuni ta’kidlash joizki, bu fan sohasi sun’iy intellekt va mashinani o‘rganishda markaziy ahamiyatini saqlab qoladi. Ma’lumotlar hajmi va murakkabligi oshgani sari, samarali xususiyatlarni yaratish va tanlashga bo‘lgan ehtiyoj yanada ortadi. Kelajakda avtomatlashtirilgan Feature Engineering usullari, ayniqsa chuqur o‘rganishga asoslangan modellar, tobora muhim rol o‘ynaydi. Ushbu yondashuvlar ma’lumotlardagi yashirin bog‘liqliklarni mustaqil ravishda aniqlashga va inson aralashuvisiz optimallashtirilgan xususiyat to‘plamlarini yaratishga qodir. Biroq, avtomatlashtirishga qaramay, soha mutaxassislarining domen bilimi va tajribasi hal qiluvchi ahamiyatga ega bo‘lib qoladi. Modelning ishonchliligini va tushunarliligini ta’minlashda insonning intuitsiyasi va ma’lumotlarni tahlil qilish mahorati ajralmasdir. Amaliy qo‘llanilishda Feature Engineeringning ahamiyati sog‘liqni saqlashdan tortib moliyagacha, ishlab chiqarishdan tortib iste’molchilar xulq-atvorini bashorat qilishgacha bo‘lgan barcha sohalarda yanada kengayadi. Axloqiy jihatlar va ma’lumotlar maxfiyligi masalalari ham Feature Engineeringni rivojlantirishda va qo‘llashda e’tiborga olinishi kerak bo‘lgan asosiy omillardir. Shuningdek, turli ma’lumotlar turlari, jumladan, strukturasiz ma’lumotlar bilan ishlash uchun yangi yondashuvlarni ishlab

chiqish ustuvor vazifa bo‘lib qoladi. Ushbu sohada doimiy tadqiqotlar va innovatsiyalar yuqori samarali va aniq mashinani o‘rganish modellarini yaratish uchun zarurdir.

Adabiyotlar, References, Литературы:

1.Bishop, C. M. Pattern Recognition and Machine Learning. Springer, 2006. – pp. 173–186.
2. Guyon, I., & Elisseeff, A. An Introduction to Variable and Feature Selection. Journal of Machine Learning Research, 2003. – pp. 1157–1182.
3. Pedregosa, F., et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 2011. – vol. 12, pp. 2825–2830.
4. Kuhn, M., & Johnson, K. Feature Engineering and Selection: A Practical Approach for Predictive Models. CRC Press, 2019. – pp. 45–78.
5. Chollet, F. Deep Learning with Python. Manning Publications, 2018. – pp. 101–115.